JMIR Publications
The Leading eHealth Publisher

In the Face of Confounders: Atrial Fibrillation Detection – Practitioners vs. ChatGPT

Yuval Avidan, Vsevolod Tabachnikov, Orel Ben Court, Razi Khoury, Amir Aker

Submitted to: Journal of Medical Internet Research
on: August 14, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

In the Face of Confounders: Atrial Fibrillation Detection – Practitioners vs. ChatGPT

Yuval Avidan[1] MD; Vsevolod Tabachnikov[1] MD; Orel Ben Court[2] MD; Razi Khoury[1] MD; Amir Aker[1] MD

[1]Carmel Medical Center, Department of Cardiology Haifa IL
[2]Carmel Medical Center, Department of Medicine Haifa IL

Corresponding Author:
Yuval Avidan MD
Carmel Medical Center, Department of Cardiology
Michal Street 7
Haifa
IL

Abstract

Background: Atrial fibrillation (AF) is the most common arrhythmia in clinical practice, yet interpretation concerns among healthcare providers persist. Confounding factors contribute to false-positive and false-negative AF diagnoses, leading to potential omissions. Artificial intelligence advancements show promise in electrocardiogram (ECG) interpretation.

Objective: We sought to examine the diagnostic accuracy of ChatGPT-4omni (GPT-4o), equipped with image evaluation capabilities, in interpreting ECGs with confounding factors and compare its performance to that of physicians.

Methods: Twenty ECG cases, divided into Group A (10 cases of AF or atrial flutter) and Group B (10 cases of sinus or another atrial rhythm), were crafted into multiple-choice questions. Total of 100 practitioners (25 from each: emergency medicine, internal medicine, primary care, and cardiology) were tasked to identify the underlying rhythm. Next, GPT-4o was prompted in five separate sessions.

Results: GPT-4o performed inadequately, averaging 3 (±2) in Group A questions and 5.40 (±1.34) in Group B questions. Upon examining the accuracy of the total ECG questions, no significant difference was found between GPT-4o, internists, and primary care physicians ($p = 0.952$ and $= 0.852$, respectively). Cardiologists outperformed other medical disciplines and GPT-4o ($p < 0.001$), while emergency physicians followed in accuracy, though comparison to GPT-4o only indicated a trend ($p = 0.068$).

Conclusions: GPT-4o demonstrated suboptimal accuracy with significant under- and over-recognition of AF in ECGs with confounding factors. Despite its potential as a supportive tool for ECG interpretation, its performance did not surpass that of medical practitioners, underscoring the continued importance of human expertise in complex diagnostics.

(JMIR Preprints 14/08/2024:65395)
DOI: https://doi.org/10.2196/preprints.65395

Original Manuscript

In the Face of Confounders: Atrial Fibrillation Detection – Practitioners vs. ChatGPT

Yuval Avidan MD [1,2], Vsevolod Tabachnikov MD [1,2], Orel Ben Court MD [2,3], Razi Khoury MD [1,2], Amir Aker MD [1,2]

1. Department of Cardiology, Lady Davis Carmel Medical Center, Haifa, Israel
2. The Ruth and Bruce Rappaport Faculty of Medicine, Technion, Israel Institute of Technology, Haifa, Israel
3. Department of Internal Medicine, Lady Davis Carmel Medical Center, Haifa, Israel

Corresponding Author:

Name: Yuval Avidan
Full Postal Mailing Address: 7 Michal St., Haifa, Israel
Telephone: +972-48250801
Fax: +972-48250802
E-mail: Yuvalavidan10@gmail.com

Wordcount: 2570

Abstract

Introduction Atrial fibrillation (AF) is the most common arrhythmia in clinical practice, yet interpretation concerns among healthcare providers persist. Confounding factors contribute to false-positive and false-negative AF diagnoses, leading to potential omissions. Artificial intelligence advancements show promise in electrocardiogram (ECG) interpretation. We sought to examine the diagnostic accuracy of ChatGPT-4omni (GPT-4o), equipped with image evaluation capabilities, in interpreting ECGs with confounding factors and compare its performance to that of physicians.

Methods Twenty ECG cases, divided into Group A (10 cases of AF or atrial flutter) and Group B (10 cases of sinus or another atrial rhythm), were crafted into multiple-choice questions. Total of 100 practitioners (25 from each: emergency medicine, internal medicine, primary care, and cardiology) were tasked to identify the underlying rhythm. Next, GPT-4o was prompted in five separate sessions.

Results GPT-4o performed inadequately, averaging 3 (±2) in Group A questions and 5.40 (±1.34) in Group B questions. Upon examining the accuracy of the total ECG questions, no significant difference was found between GPT-4o, internists, and primary care physicians (p = 0.952 and = 0.852, respectively). Cardiologists outperformed other medical disciplines and GPT-4o (p < 0.001), while emergency physicians followed in accuracy, though comparison to GPT-4o only indicated a trend (p = 0.068).

Conclusion GPT-4o demonstrated suboptimal accuracy with significant under- and over-recognition of AF in ECGs with confounding factors. Despite its potential as a supportive tool for ECG interpretation, its performance did not surpass that of medical practitioners, underscoring the continued importance of human expertise in complex diagnostics.

Keywords: Electrocardiogram, Chat-GPT, Physicians, Atrial fibrillation

Introduction

Atrial fibrillation (AF) is the predominant arrhythmia in clinical practice [1]. Despite the fact that proficiency in interpreting electrocardiograms (ECGs) is essential for medical practitioners [2], major concerns exist about the interpretation skills of healthcare providers. Several studies indicate suboptimal levels among internists, emergency physicians [3,4], primary care physicians [5], and even cardiologists [6]. Numerous confounding factors contributing to both false-positive and false-negative diagnoses of AF have been identified [7–9], complicating the diagnostic process and potentially leading to significant medical omissions.

Artificial intelligence (AI) is transforming the field of medicine by enabling machines to perform tasks that historically relied on human intelligence [10]. One of the most prominent models is OpenAI's ChatGPT (Generative Pre-trained Transformer), has evolved significantly since its initial introduction. Several iterations have excelled in medical examinations, outperforming cardiologists and emergency physicians in the interpretation of ECGs [11–13]. Released on May 13, 2024, the ChatGPT-4omni (GPT-4o), exceeds its predecessors with improved multimodal capabilities. Unlike previous versions, GPT-4o image analysis feature lets users upload and analyze medical images, incorporating the results into its decision-making.

The aforementioned underlines the clinical relevance of developing an integrated AI tool to aid healthcare providers in interpreting ECGs for AF, aiming to reduce diagnostic errors, particularly when encountering ECGs complicated by confounding factors. The competency of the image analysis feature of GPT-4o has not been previously studied. We aimed to evaluate the diagnostic accuracy of GPT-4o in interpreting images of ECG tracings in the presence of confounding factors, and to compare its performance with that of medical practitioners.

Material and Methods

This proof-of-concept study evaluated the customized GPT model "ECG-Analyst" for its ability to identify underlying atrial rhythms in ECGs complicated by confounding factors. The collection and validation of ECG tracings were conducted as follows: two cardiologists (Y.A, I.N) performed a web search, yielding 33 eligible ECGs from various sources with permission for reuse specifically for this investigation. Two consultant cardiologists (V.T, R.K), blinded to the original interpretations, independently provided their diagnoses for each ECG. In cases of disagreement, a third consultant (A.A) served as the arbitrator. Quality assessment led to the exclusion of 13 ECGs due to poor image quality or incomplete data, resulting in a final set of 20 ECGs formulated as multiple-choice questions. The tracings were divided into two groups: Group A, consisting of 10 ECGs with AF or AFL, and Group B, comprising 10 cases of non-AF/AFL ECGs. Each tracing was accompanied by a standardized clinical history: "an 80-year-old man with weakness for several days".

The study encompassed a total of 100 medical practitioners: 25 emergency medicine physicians, 25 internists, 25 primary care physicians, and 25 cardiologists, each with fewer than five years of clinical experience. Participants were tasked with identifying whether the underlying rhythm depicted in each ECG was AF, AFL, or non-AF/AFL [e.g. sinus rhythm, ectopic atrial rhythm, multifocal atrial tachycardia (MAT), supraventricular tachycardia, atrial pacing, ventricular rhythm, etc.]. Each question presented three potential answers, of which only one was correct. The detailed questions are outlined in Table 1. For example, if the correct answer was AF, then selecting AFL was considered incorrect. The questions were administered using Google Forms and time restriction was 15-minute. All physicians were contacted via their hospital-affiliated E-Mail addresses and completed the questionnaire anonymously. Prior to commencement, participants were briefed on the study's scope and objectives.

In the subsequent phase, the same 20 questions were presented to the GPT-4o model equipped with the image evaluation feature, prompting it to provide "the most suitable answer”. To account for reproducibility, the questions were presented 5 times over consecutive days, with the order of the questions varied each day, and the responses were documented. The answers from the emergency medicine physicians, internists, primary care physicians, cardiologists, and Chat-GPT were examined separately across three categories: overall success rate, missed diagnosis of AF or AFL, and overdiagnosis of AF or AFL. The study was approved by the Review Board of the Carmel Medical Center and adhered to the principles of the Declaration of Helsinki.

Power analysis

With a significance level set at 0.05, power at 0.95, and large effect size at 0.8 (Cohen's d), the sample size was calculated using the G Power program to be 25 participants for each group, totaling 100 participants.

Statistical analysis

To test the differences in total scores (groups A and B) among physicians from the different fields, a one-way analysis of variance (ANOVA) was conducted. The Shapiro-Wilk test was used to assess the normality of total scores, and Levene's test evaluated the equality of variances across groups. A Tukey post-hoc test was performed to identify specific group differences. To evaluate the differences in Group A and in B scores between groups of doctors, the Shapiro-Wilk test assessed the normality of scores. Some groups did not conform to normality, so the Kruskal-Wallis test was used for comparisons. A post-hoc test examined specific group differences. Non-normal distribution in some groups led to the use of the Kruskal-Wallis test for comparisons. Statistical analysis was performed with IBM SPSS Statistics 25, and P values less than 0.05 were considered statistically significant.

Internal medicine	Primary care	Emergency medicine	GPT-4o	Cardiology		
Group A: AF/AFL						
16	(12*) 8	(12*) 32	60	60	AF in the presence of ventricular paced rhythm and low-amplitude atrial activity [14]	1
(28*) 56	(48*) 40	(20*) 72	(40*) 60	(8*) 84	AFL with variable AV conduction [15]	2
(8*) 44	(12*) 16	(8*) 60	(20*) 40	(4*) 68	Pre-excited AF in WPW [16]	3
56	(8*) 20	(16*) 64	(40*) 40	(20*) 72	AF with rapid ventricular rate and transient RBBB (Ashman phenomenon) [17]	4
20	20	28	60	60	AF with rapid ventricular rate and extremely wide LBBB [18]	5
52	56	52	0	84	AFL 2:1 conduction (Fig. 1) [19]	6
52	8	36	40	72	AF with biventricular pacing and PVCs [20]	7
32	24	52	0	72	AF with RVH and low amplitude atrial activity [21]	8
(16*) 36	(12*) 36	(28*) 56	0	(8*) 80	AF with wide QRS [22]	9
(12*) 64	(16*) 52	76	0	96	AF and low voltage QRS [23]	10
Group B: non-AF/AFL						
20	32	44	60	56	Multifocal atrial tachycardia [24]	11
40	20	52	0	68	Sinus with baseline artifact mimicking AFL [25]	12
68	52	75	60	96	Sinus bradycardia, first degree AV block and PVCs [25]	13
44	28	48	80	60	Sinus with PACs and low calibration of 5 mm/mV *(Fig. 2)*	14
88	68	80	100	96	Sinus with atrial couplets [26]	15
44	24	44	0	72	Sinus with PACs, prolonged PR interval and wide QRS [25]	16
48	60	68	60	88	Sinus with PACs and poor baseline quality [27]	17
60	48	76	0	96	Sinus with baseline artifact [28]	18

56	56	68	100	92	Sinus with U-wave and low P-wave amplitude [29]	19
24	20	52	80	68	Multifocal atrial tachycardia [30]	20

Table 1: Questions and the percentage of correct responses according to different subgroups. **AF**: Atrial Fibrillation **AFL**: Atrial Flutter **AV**: Atrioventricular **ECG**: Electrocardiogram **LBBB**: Left Bundle Branch Block **PACs**: Premature Atrial Contractions **PVCs**: Premature Ventricular Contractions **RBBB**: Right Bundle Branch Block **RVH**: Right Ventricular Hypertrophy **WPW**: Wolff-Parkinson-White. * The percentage of participants who misclassified AFL as AF, or vice versa, in Group A questions.

Results

Significant differences were observed among the different participant groups in both the AF/AFL questions, non-AF/AFL ECG questions, and total ECG questions (P < 0.001). GPT-4o scored only 3 (±2) out of the 10 Group A questions, and an average of 5.40 (±1.34) in Group B, the non-AF/AFL questions (P < 0.001). Cardiologists outperformed the other groups by providing an average score of 7.48 (±1.38) correct answers in AF/AFL ECG questions, an average of 7.92 (±1.46) correct answers in the non-AF/AFL ECG questions, and an average of 15.4 (±2.21) correct answers in total ECG questions. The average score of the other groups was 8.4 (±2.50), 11.36 (±2.36), 6.88 (±2.36), 9.2 (±2.16), for GPT-4o, emergency medicine physicians, internists, and primary care physicians, respectively (P < 0.001). The lowest average scores were documented among primary care physicians, with an average of 2.8 (±1.41) in Group A questions, and 4.08 (±1.80) in Group B ECG questions (P < 0.001). Detailed information regarding the data is presented in Table 2.

		(%) Mean	SD	N	P value
Total	Cardiology	(77) 15.4	2.21	25	0.001>
	GPT-4o	(42) 8.4	2.50	5	
	Emergency medicine	(56.8) 11.36	2.36	25	
	Primary care	(34.4) 6.88	2.31	25	
	Internal medicine	(46) 9.2	2.16	25	
Group A: AF/AFL	Cardiology	(74.8) 7.48	1.38	25	0.001>
	GPT-4o	(30) 3.0	2	5	
	Emergency medicine	(52.8) 5.28	1.13	25	

	Primary care	(28) 2.80	1.41	25	
	Internal medicine	(42.8) 4.28	1.36	25	
Group :B Non-AF/AFL	Cardiology	(79.2) 7.92	1.46	25	0.001>
	GPT-4o	(54) 5.40	1.34	5	
	Emergency medicine	(60.8) 6.08	1.63	25	
	Primary care	(40.8) 4.08	1.80	25	
	Internal medicine	(49.2) 4.92	1.25	25	

Table 2: Evaluation of the number of correct answers according to groups and categories.

Upon examining the relationship between the groups; in Group A, the AF/AFL ECG questions, no significant difference was detected between GPT-4o, internists, primary care physicians, and emergency medicine physicians (p = 1.000, p = 1.000, p = 0.329). Cardiologists were found to be more successful than compared to the emergency physicians (p = 0.008) and other subgroups (p < 0.001). In Group B, the non-AF/AFL ECG questions, cardiologists again were observed to be more successful compared to internists, primary care physicians, and emergency medicine physicians (p < 0.001, p < 0.001, p = 0.03), however, no significant difference was found when compared to GPT-4o (p = 0.17). Among the other subgroups, no statistical difference was observed between the interpretations of GPT-4o, internists, primary care physicians, and emergency medicine physicians (p = 1.000). Lastly, when the accuracies of total ECG questions were examined, cardiologists demonstrated the highest accuracy in ECG interpretation, surpassing both other doctors and GPT-4o (p < 0.001). Emergency physicians followed in accuracy, outperforming internists and primary care physicians (p = 0.01, p < 0.001), whereas comparison to GPT-4o only indicated a trend rather than statistical significance (p = 0.068). GPT-4o was found to perform similarly to both internists and primary care physicians (p = 0.952, p = 0.852). More detailed information is presented in Table 3.

	Groups comparison	P value
Group A: AF/AFL	Cardiology – GPT-4o	*0.001> **
	Cardiology – Emergency medicine	**0.008
	Cardiology - Primary care	*0.001> **
	Cardiology – Internal medicine	*0.001> **

	GPT-4o – Emergency medicine	0.329
	GPT-4o - Primary care	1
	GPT-4o -Internal medicine	1
	Emergency medicine - Primary care	*0.001>**
	Emergency medicine - Internal medicine	0.623
	Primary care - Internal medicine	0.117
:Group B Non-AF/ AFL	Cardiology - GPT-4o	0.17
	Cardiology - Emergency medicine	*0.03
	Cardiology - Primary care	**001.>*
	Cardiology - Internal medicine	**001.>*
	GPT-4o - Emergency medicine	1
	GPT-4o - Primary care	1
	GPT-4o - Internal medicine	1
	Emergency medicine - Primary care	**0.005
	Emergency medicine - Internal medicine	0.282
	Primary care - Internal medicine	1
Total ECGs	Cardiology - GPT-4o	*0.001>**
	Cardiology - Emergency medicine	*0.001>**
	Cardiology - Primary care	*0.001>**
	Cardiology - Internal medicine	*0.001>**
	GPT-4o - Emergency medicine	0.068
	GPT-4o - Primary care	0.852
	GPT-4o - Internal medicine	0.952
	Emergency medicine - Primary care	*0.001>**

	Emergency medicine - Internal medicine	*0.01
	Primary care - Internal medicine	**0.004

Table 3: The Relationship Between the Groups. *p < 0.05, **p < 0.01, **p < 0.001

According to our study results, GPT-4o consistently answered questions 6, 8, 9, 10, 12, 16, and 18 incorrectly. GPT-4o struggled particularly with questions 8, 9, and 10, which all displayed low fibrillatory wave amplitude (FWA), and questions 12 and 18, which were complicated by baseline artifacts. Conversely, GPT-4o received the highest average scores on five questions: 11, 14, 15, 19, and 20. Specifically, questions 11 and 20 depicted MAT; question 14 displayed sinus rhythm with premature atrial contractions (PACs) and low calibration of 5 mm/mV; and questions 15 and 19 depicted atrial couplets and U-waves, respectively. Detailed information regarding the distribution of correct answers per question is presented in Table 1.

Discussion

Our primary aim was to evaluate the diagnostic accuracy of the newly developed image analysis feature of GPT-4o in interpreting ECGs complicated by confounding factors. By comparing its performance to that of medical practitioners, our findings underscore the limitations of the customized GPT model in accurately identifying AF under these challenging conditions.

Our study elucidates several key observations. Firstly, GPT-4o demonstrated a level of susceptibility to electrocardiographic confounders similar to that of human interpreters, which affected its diagnostic accuracy. Many confounding factors contributed to false negative interpretation of AF, including low FWA, pre-excited AF, Ashman phenomenon, paced-rhythms, PVCs, AFL with variable atrioventricular (AV) conduction, and very rapid or slow ventricular rates. In particularly, the presence of paced-rhythm negatively influenced the diagnostic accuracy across all fields. This aligns with existing literature reporting low accuracy rates among hospital physicians, primary care physicians, and even board-certified cardiologists [8,14,31]. Intriguingly, with computer-based ECG interpretation, by far, the most common errors are related to paced rhythms [8]. Clinically, this is relevant due to the high incidence of AF in pacemaker patients, which is exacerbated by chronic right ventricular pacing [32]. Considering the low FWA depicted in questions 8, 9, and 10, it is evident that unlike coarse AF, fine AF exhibits oscillations with amplitudes less than 0.1 mV, making it more prone to elude detection. While medical practitioners, especially cardiologists and emergency physicians, performed at an above satisfactory level, GPT-4o repeatedly responded incorrectly during all prompts, falsely categorizing these tracings as sinus rhythm. We asked it to explain why it struggles with the recognition of AF in this context, and the response was as follows: *“These subtle* *oscillations can be mistaken for baseline noise or sinus rhythm. Unlike human cardiologists who are trained to recognize these faint signals, the AI model may lack the nuanced understanding and contextual knowledge required to accurately identify fine AF".*

Noteworthy, a missed diagnosis of pre-excited AF can be particularly detrimental. If this unique pattern is misinterpreted as a ventricular arrhythmia and treated with amiodarone, it can paradoxically lead to ventricular fibrillation [33]. According to our data, the diagnostic accuracy in question 3 was highest among cardiologists (68%) but unacceptably insufficient across other groups, with notably low rates among internists (44%), GPT-4o (40%), and primary care physicians (16%). Additionally, as reflected by question number 2, both GPT-4o and a substantial number of doctors misclassified AFL as AF (48% of primary care physicians, 28% of internists, and GPT-4o in 2 out of 5 prompts). Not unfrequently, AFL with variable AV conduction is misinterpreted as AF due to the irregularity [34], which unlike cases of AF, is caused by the changing refractoriness of the AV node. This highlights a common misconception about the clinical relevance of distinguishing between these two arrhythmias. While they share similarities in management, their response to pharmacotherapy and rhythm control strategies differs, as reflected in current guidelines [35]. Moreover, as demonstrated by tracing number 6, diagnostic challenges emerge when AFL presents with a 2:1 conduction ratio. Similar to many computerized ECG algorithms, GPT-4o consistently failed to identify 2:1 AFL (Figure 1.) [36]. In this situation, the underlying atrial depolarization is partially masked, as the flutter waves are 'buried' within the T-wave or QRS complexes, which makes discrimination difficult for computer algorithms.

As our findings indicate, several confounding factors could contribute to the false positive diagnosis of AF in both physicians and GPT-4o interpretations. Consistent with findings from previous studies [8,37,38], factors such as PACs, reduced ECG quality, baseline artifacts, low-amplitude P-waves, and supraventricular tachycardias, can all lead to false-positive diagnoses of AF. Analyzing GPT-4o's responses to the ECGs in Group B, its overall performance was deemed average. Despite some correct interpretations, the model's performance was particularly compromised in cases number 12 and 18, both involving movement artifacts that mimic fibrillatory waves, resulting in incorrect responses that the majority of cardiologists and emergency medicine specialists were able to avoid.

In several tracings, GPT-4 demonstrated the highest accuracy rates, with one particularly notable diagnosis being MAT, as depicted in questions 11 and 20. As shown in Table 1, medical practitioners have performed poorly in identifying MAT, a relatively common arrythmia. Historically, MAT has been notoriously labeled as a frequent arrhythmia misclassified as atrial fibrillation by many physicians and ECG computer analysis programs [39]. GPT-4o accurately explained its analytical decision-making process by noting the presence of an irregular rhythm, a rapid atrial rate, at least three different P wave morphologies, and variable P-P, P-R, and R-R intervals. The distinction between MAT and AF is crucial, as the treatment of MAT primarily focuses on managing underlying conditions, such as heart failure or pulmonary disease exacerbation, rather than adhering to strict rate or rhythm control strategies commonly employed in AF. Moreover, unlike AF, anticoagulation therapy plays no role in the management of MAT.

Our data contradict the findings of a recent study, which reported that an earlier iteration, GPT-4, outperformed cardiologists and emergency medicine specialists in analyzing routine ECG questions and was non-inferior to cardiologists in complex cases [11]. However, this notion should be taken with a grain of salt due to major methodological gaps in the study design, particularly the use of ECG descriptions rather than actual ECG images. Therefore, while the textual interpretation of ECGs may offer potential educational benefits, asserting its research or clinical relevance seems questionable. Besides, as depicted in Table 1, another major pitfall in utilizing GPT-4o is its relative inconsistency when prompted with the same query, despite unchanged structure and content. These variations may be attributed to the model's underlying structure, learning architecture, and parameter settings, making it challenging to reproduce consistent responses to repeated queries.

Finally, while not the primary focus of this study, our findings uncovered significant knowledge gaps and alarmingly low accuracy rates among certain subgroups, specifically primary care physicians (40.8%) and internists (49.2%). This resulted in substantial over- and under-diagnosis of AF and AFL in the presence of confounding factors. These results align with previously published real-world outcomes [5,40], and raise concerns about the adequacy of post-graduate ECG education, underscoring the pressing need to address this gap.

To conclude, although GPT-4o model showed some correct interpretations, its overall performance was not superior to that of medical practitioners, with unacceptably high rate of under- and over-recognition of AF and AFL in ECG tracings complicated by confounding factors. This finding is crucial, as it highlights that despite the theoretical appeal of AI advancements in clinical settings, human expertise remains indispensable. The superiority of cardiologists, and to some extent emergency physicians, is particularly evident in complex diagnostic scenarios. Similar to non-expert ECG readers, the chatbot exhibited a pronounced susceptibility to electrocardiographic confounders. Despite its potential as a supportive tool for ECG interpretation, our experience indicates that substantial development in GPT-4o's diagnostic capabilities is warranted before it can be reliably integrated into clinical practice. Nevertheless, it is too early to render a final verdict on this emerging innovative technology. Future research should focus on refining these technologies to better address the complexities of clinical diagnostics.

Limitations

Our study did not aim to comprehensively evaluate GPT-4o's overall proficiency in ECG interpretation. Instead, it focused on assessing the model's performance in interpreting ECGs complicated by confounding factors, particularly in the context of AF and AFL. Hence, further research is needed before drawing firm conclusions about other aspects of GPT-4o’s diagnostic capabilities in general ECG interpretation.

Disclosures: The authors have nothing to disclose.

Funding: None.

Contributions: All co-authors have substantially contributed to the manuscript. Y.A contributed in review & editing, supervision, methodology, conceptualization. O.B, R.K and V.T contributed in data curation and writing the original draft. A.A contributed with formal analysis, supervision, conceptualization, review and editing

Data availability: All data are available upon reasonable request

References

1. Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, Blomström-Lundqvist C, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J 2021;42(5):373–498.

2. Salerno SM, Alguire PC, Waxman HS. Training and Competency Evaluation for Interpretation of 12-Lead Electrocardiograms: Recommendations from the American College of Physicians*. Ann Intern Med 2003;138(9):747.

3. Berger JS, Eisen L, Nozad V, D’Angelo J, Calderon Y, Brown DL, et al. Competency in electrocardiogram interpretation among internal medicine and emergency medicine residents. Am J Med 2005;118(8):873–80.

4. de Jager J, Wallis L, Maritz D. ECG interpretation skills of South African Emergency Medicine residents. Int J Emerg Med 2010;3(4):309–14.

5. Mant J, Fitzmaurice DA, Hobbs FDR, Jowett S, Murray ET, Holder R, et al. Accuracy of diagnosing atrial fibrillation on electrocardiogram by primary care practitioners and interpretative diagnostic software: analysis of data from screening for atrial fibrillation in the elderly (SAFE) trial. BMJ 2007;335(7616):380.

6. Cook DA, Oh SY, Pusic M V. Accuracy of Physicians' Electrocardiogram Interpretations. JAMA Intern Med 2020;180(11):1461.

7. Alkhalil M, Prabhavalkar S, Cromie N. Atrial fibrillation in ventricular-paced rhythm: under-recognised, underdiagnosed and potentially dangerous. Heart Asia 2014;6(1):39–40.

8. Davidenko JM, Snyder LS. Causes of errors in the electrocardiographic diagnosis of atrial fibrillation by physicians. J Electrocardiol 2007;40(5):450–6.

9. Bogun F, Anh D, Kalahasty G, Wissner E, Bou Serhal C, Bazzi R, et al. Misdiagnosis of atrial fibrillation and its clinical consequences. Am J Med 2004;117(9):636–42.

10. McGrow K. Artificial intelligence. Nursing (Brux) 2019;49(9):46–9.

11. Günay S, Öztürk A, Özerol H, Yiğit Y, Erenler AK. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. Am J Emerg Med 2024;80:51–60.

12. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? European Heart Journal - Digital Health 2023;4(3):279–81.

13. Knoedler L, Alfertshofer M, Knoedler S, Hoch CC, Funk PF, Cotofana S, et al. Pure Wisdom or Potemkin Villages? A Comparison of ChatGPT 3.5 and ChatGPT 4 on USMLE Step 3 Style Questions: Quantitative Analysis. JMIR Med Educ 2024;10:e51148.

14. Alkhalil M, Prabhavalkar S, Cromie N. Atrial fibrillation in ventricular-paced rhythm: under-recognised, underdiagnosed and potentially dangerous. Heart Asia 2014;6(1):39–40.

15. Grauer K. ECG Basics: Atrial Flutter With Variable Conduction [Internet]. 2015 [cited 2024 Jun 26];Available from: https://www.ecgguru.com/ecg/atrial-flutter-variable-conduction

16. Nathanson LA MSSCG AL. ECG Wave-Maven: Self-Assessment Program for Students and Clinicians, Case Number 15 [Internet]. 2024 [cited 2024 Jun 26];Available from: https://ecg.bidmc.harvard.edu/mavendata/images/case15/img.pdf

17. Nathanson LA MSSCGAL. ECG Wave-Maven: Self-Assessment Program for Students and Clinicians, Case Number 88 [Internet]. 2024 [cited 2024 Jun 26];Available from: https://ecg.bidmc.harvard.edu/mavendata/images/case88/img.pdf

18. Nathanson LA MSSCGAL. ECG Wave-Maven: Self-Assessment Program for Students and Clinicians, Case Number 111 [Internet]. 2024 [cited 2024 Jun 26];Available from: https://ecg.bidmc.harvard.edu/mavendata/images/case111/img.pdf

19. McLaren J. Emergency Medicine Cases: Approach to Atrial Fibrillation [Internet]. 2023 [cited 2024 Jul 19];Available from: https://emergencymedicinecases.com/ecg-cases-approach-atrial-fibrillation/

20. Dewaswala N. Determining pacemaker type from EKG [Internet]. 2021 [cited 2024 Jul 20];Available from: https://www.medicowesome.com/2021/11/determining-pacemaker-type-from-ekg-rv.html

21. Nathanson LA MSSCG AL. ECG Wave-Maven: Self-Assessment Program for Students and Clinicians [Internet]. 2024 [cited 2024 Jul 20];Available from: https://ecg.bidmc.harvard.edu/mavendata/images/case18/img.pdf

22. Nathanson LA MSSCG AL. ECG Wave-Maven: Self-Assessment Program for Students and Clinicians [Internet]. 2024 [cited 2024 Jul 20];Available from: https://ecg.bidmc.harvard.edu/mavendata/images/case127/img.pdf

23. Nathanson LA MSSCG AL. ECG Wave-Maven: Self-Assessment Program for Students and Clinicians [Internet]. 2024 [cited 2024 Jul 21];Available from: https://ecg.bidmc.harvard.edu/mavendata/images/case201/img.pdf

24. Multifocal Atrial Tachycardia (MAT). (2024). https://litfl.com/multifocal-atrial-tachycardia-mat-ecg-library/. Accessed: 22 June, 2024:

25. Lindow T, Kron J, Thulesius H, Ljungström E, Pahlm O. Erroneous computer-based interpretations of atrial fibrillation and atrial flutter in a Swedish primary health care setting. Scand J Prim Health Care 2019;37(4):426–33.

26. Nathanson LA MSSCG AL. ECG Wave-Maven: Self-Assessment Program for Students and Clinicians, Case Number 418 [Internet]. 2024 [cited 2024 Jun 27];Available from: https://ecg.bidmc.harvard.edu/mavendata/images/case418/img.pdf

27. Bogun F, Anh D, Kalahasty G, Wissner E, Bou Serhal C, Bazzi R, et al. Misdiagnosis of atrial fibrillation and its clinical consequences. Am J Med 2004;117(9):636–42.

28. Mond H. Fact or Artefact [Internet]. 2021 [cited 2024 Jul 7];Available from: https://singapore.cardioscan.co/blog/resource/fact-or-artefact/

29. Nathanson LA MSSCG AL. ECG Wave-Maven: Self-Assessment Program for Students and Clinicians [Internet]. 2024 [cited 2024 Jul 21];Available from: https://ecg.bidmc.harvard.edu/mavendata/images/case189/img.pdf

30. Nathanson LA MSSCG AL. ECG Wave-Maven: Self-Assessment Program for Students and Clinicians [Internet]. 2024 [cited 2024 Jul 21];Available from: https://ecg.bidmc.harvard.edu/maven/dispcase.asp?rownum=363&ans=0&caseid=364

31. Patel AM, Westveer DC, Man KC, Stewart JR, Frumin HI. Treatment of underlying atrial fibrillation: paced rhythm obscures recognition. J Am Coll Cardiol 2000;36(3):784–7.

32. Costa MAC da, Santos JF do LP dos, Schafranski MD. Prevalence of Atrial Fibrillation in Pacemaker Patients. International Journal of Cardiovascular Sciences 2022;

33. Brugada J, Katritsis DG, Arbelo E, Arribas F, Bax JJ, Blomström-Lundqvist C, et al. 2019 ESC Guidelines for the management of patients with supraventricular tachycardiaThe Task Force for the management of patients with supraventricular tachycardia of the European Society of Cardiology (ESC). Eur Heart J 2020;41(5):655–720.

34. Shiyovich A, Wolak A, Yacobovich L, Grosbard A, Katz A. Accuracy of Diagnosing Atrial Flutter and Atrial Fibrillation From a Surface Electrocardiogram by Hospital Physicians: Analysis of Data

From Internal Medicine Departments. Am J Med Sci 2010;340(4):271–5.

35. Hindricks G, Potpara T, Dagres N, Arbelo E, Bax JJ, Blomström-Lundqvist C, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). Eur Heart J 2021;42(5):373–498.

36. Bui QA, Farrell R, Rigales L, Hoffman I. Why Do Computer Programs Misdiagnose Flutter? Am J Med 2016;129(11):e289–90.

37. Wang Y, Seow S, Singh D, Poh K, Chai P. Rhythmic chaos: irregularities of computer ECG diagnosis. Singapore Med J 2017;58(9):516–20.

38. Littmann L. Common ECG interpretation software mistakes Part III: Computer errors that should never be missed. J Electrocardiol 2023;81:281–4.

39. Varriale P, David W, Chryssos BE. Multifocal atrial arrhythmia—A frequent misdiagnosis? A correlative study using the computerized ECG. Clin Cardiol 1992;15(5):343–6.

40. Berger JS, Eisen L, Nozad V, D’Angelo J, Calderon Y, Brown DL, et al. Competency in electrocardiogram interpretation among internal medicine and emergency medicine residents. Am J Med 2005;118(8):873–80.

Supplementary Files